2 Speaking for listening

Anne Cutler

Abstract

Speech production is constrained at all levels by the demands of speech perception. The speaker's primary aim is successful communication, and to this end semantic, syntactic and lexical choices are directed by the needs of the listener. Even at the articulatory level, some aspects of production appear to be perceptually constrained, for example the blocking of phonological distortions under certain conditions. An apparent exception to this pattern is word boundary information, which ought to be extremely useful to listeners, but which is not reliably coded in speech. It is argued that the solution to this apparent problem lies in rethinking the concept of the boundary of the lexical access unit. Speech rhythm provides clear information about the location of stressed syllables, and listeners do make use of this information. If stressed syllables can serve as the determinants of word lexical access codes, then once again speakers are providing precisely the necessary form of speech information to facilitate perception.

Introduction

The central argument of this chapter is that speech production is subject to perceptual constraints. Since speakers speak chiefly to communicate with listeners, it might seem quite unremarkable to claim that speakers construct their speech output so as to cater for listeners' needs. However, perceptually determined constraints on production turn out to be remarkably pervasive in the production process. Even at quite 'low' levels, i.e. relatively close to output, the production of an utterance is constrained by factors which have more to do with the nature of the listeners' perceptual process than with the nature of the production process itself. This chapter

LANGUAGE PERCEPTION AND PRODUCTION
ISBN 0-12-052750-2

will summarize evidence on the production of nonce-words, on the correction of slips of the tongue, and on the application of optional phoneme elision and assimilation rules, all of which shows sensitivity to perceptual factors constraining the production process. In the final sections it will be shown that even the task of word boundary detection, which is one of the chief difficulties of speech perception, appears to be facilitated by certain aspects of speech production: speakers provide rhythmic cues on which listeners can base a strategy of segmentation.

Speech as communication

A speaker's primary aim is to formulate a *message.* Thus what speakers say is, in most cases, what they want listeners to hear — not just what they want to utter to satisfy some purely internal articulatory need which could as easily be satisfied without a listener. The content of the message, that is, is determined by (the speaker's perception of) the characteristics of the listener. Likewise, listener characteristics can determine aspects of the message's form — speakers speak more simply to children, for instance, and to people with an imperfect grasp of the language in question. Some syntactic constructions are harder to process than others, as much psycholinguistic research has demonstrated; speakers replace harder constructions with easier ones when communication seems to be unsuccessful (Valian and Wales, 1976). Formal versus informal registers (with their consequent syntactic and lexical elaboration versus simplification) are chosen on the basis of the current relationship between speaker and listener. Speakers draw on their knowledge of what their listeners already know in choosing what to say and how exactly to express it. Consider Grice's four maxims of conversation: be brief, be relevant, be polite, be sincere. Rephrased, they exhort speakers to avoid boring, offending or deceiving their listeners.

If speech is to function effectively as the performance of a communicative act, the speaker must obviously cater to the listener's needs. Perhaps slightly less predictable, however, is the degree to which actual lexical selection can be subject to influences arising from the nature of the perception process. Choice of speech register, mentioned above, can of course have implications for lexical selection — determining a high frequency word rather than its low frequency synonym, choosing between a specific versus a more general term, using or avoiding the taboo adjective. But quite general perceptual constraints seem to apply even to lexical processes which do not involve social considerations. A case in point is the way speakers fill a momentarily felt lexical gap by making up a novel word, as described in the next section.

Perceptual constraints on word creation

When there is a choice between alternative word formations — e.g. for making a nonce verb out of a noun — the version which produces the more easily perceptible result is consistently preferred. In a series of experiments comparing speakers' preferences for different types of suffix (Cutler 1980a, 1981), for example, I found that suffixes which attach to an existing word without affecting its phonological structure (e.g. *-ness, -ish)* were chosen in preference to suffixes which resulted in a change in the base word's stress pattern or vowel quality. Thus in Table 1, the neologisms on the left in the upper portion of the table (all of which preserve without alteration the word to which they have attached) were preferred to those on the right, in which the original word is less perfectly preserved. *Incestuousness,* for instance, contains *incestuous* unaltered within it, whereas *incestuosity* changes the final vowel of *incestuous* and also shifts the primary stress from the second to the fourth syllable.

Table 1. Neologism preferences.

incestuousness	>	incestuosity
dowagerish	>	dowagerial
ambiguize	>	ambiguify
comprisement	>	comprision
but		
jejuneness	=	jejunity
auctioneerish	=	auctioneerial
splendidise	=	splendify
excusement	=	excusion

These results do not just reflect preference for some particular suffixes (e.g. *-ness)* over others (e.g. *-ity);* it is only whether or not the original form remains intact in the neologism which matters. This is clear from the further finding that no preference is shown between two suffix types if neither of them alters the phonology of the existing word to which they are attached. The pairs of neologisms in the bottom half of Table 1 exemplify this comparison. *Jejune,* for instance, is equally well preserved in *jejuneness* and *jejunity; auctioneer* in *auctioneerish* and *auctioneerial.* Speakers show equal preference for both members of such pairs.

The best neologism, then, is one in which the word on which it is based is transparently preserved. Neologisms, by definition, do not have entries in language users' mental lexicons; so the task of understanding a neologism must differ from the usual task of word identification. It can only be done, in fact, by dividing the neologistic affix from the rest of the form and processing the two parts separately.[1] A neologism which contains an intact known word, therefore, will easily separate into a lexically accessible word plus suffix; understanding it will be a simple matter of recognizing the existing word and combining it with the new (perhaps syntactically transforming) suffix to derive the novel meaning. If the known word is not transparently preserved, however, understanding the neologism will necessitate further procedures of undoing the phonological transformations before any available lexical entry can be accessed. The neologism preference data therefore seem to indicate that the determining factor in nonce word formation is how easily the new creation can be understood; i.e. speakers create neologisms with the listeners' needs in mind.

If simple transparency were the only issue in these results, however, it might be possible to construct an argument that transparent neologisms are preferred because they are easier to *produce* rather than because they are easier to *understand:* producing a new form involves using an existing lexical entry and adding a suffix to it, and it might be claimed that it is just easier to do this if an already available articulatory programme can simply be compiled with no alteration other than an added appendage in the form of a suffix. But the results of the acceptability experiments indicate that it is not necessary to preserve the entire known form for the new derived form to be functionally transparent; it is vitally important only that the initial portions of the known word be preserved intact, but the final segments may actually be distorted. Consider the last two examples in Table 1: *splendid* has lost its final segment in *splendify,* yet *splendify* is considered to be no less acceptable than *splendidise;* likewise *excusement* and *excusion* were rated equally acceptable, although only the first preserves the final [z] sound of the verb *excuse.* In each case it appears to be sufficient to preserve the first six phonetic segments of the known word — and as it happens in each case there is no other word in the English language beginning with those six segments. All words beginning [splend] are part of the morphological family which includes *splendid;* only the verb and noun reading of *excuse* begin with [ikskju]. Thus what appears to be important in neologism formation is preserving just enough of a known word, going from left to right, to distinguish it from other words of the language — strong evidence that the transparency criterion is perceptual rather than productive.

The limits of perceptual constraints on production

Speakers cater to listeners' needs not only at the highest levels of the speech production process, such as message formulation and choice of style, but also at the lexical selection stage, even in the creation of novel word forms — as the previous section showed, a primary factor in this process is ensuring perceptibility. Yet we might expect that there would be limits on the degree to which perceptual constraints could operate in speech production. For instance, at the articulatory output level, it is reasonable to assume that it would be rather surprising to find perceptual constraints in operation. The way in which an articulatory programme is realized in muscle commands, jaw and tongue movements, and so on, is presumably not dependent on anything other than physiological factors concerned with the speaker's articulatory apparatus, plus accidental effects of the articulatory environment (e.g. the constraints imposed by trying to speak underwater, or with a mouth full of food, etc.). The actual execution of the motor programme once begun can therefore be considered immune from perceptual effects. However, it will be argued here that every conceptually prior level of speech production — that is, every level of the process up to and including compilation of the programme for articulation — is subject to constraints which derive ultimately from the communicative function of speech, and the constraints which the nature of the speech perception process places upon the successful realization of this function.

The preceding sections outlined some uncontroversial ways in which listener constraints affect higher levels of the production process, and some less obvious effects of perceptual factors upon lexical processes including augmentation of the lexical stock. The following sections concentrate upon levels of the production process which are rather closer to output — that is, those levels between word selection and articulation. In those intervening levels choices are made which will eventually constrain the details of the articulatory programme — i.e. precisely how the chosen words are to be uttered. One of the variables which can be manipulated at this level is prosodic structure; another is clarity of articulation of individual segments. Each of these will be briefly addressed in the immediately following sections, which summarize some cases in which speakers' articulatory choices appear to be rather surprisingly constrained by factors to do with the listener. The succeeding sections will then consider a case where it might seem that, again rather surprisingly, the listener is being denied assistance which the speaker could easily afford.

Some uses of accent

In the most general sense, the role of prosody in the production of an utterance is to assist in the communicative function — the speaker uses prosody to direct and control aspects of the listener's perception. Sentence accent, for example, is used to highlight new information (that is, information which the speaker believes to be new to the listener, not information which is new to the speaker). Although some linguists have devoted considerable effort to describing sentence accent placement in terms of syntactic structure, such descriptions are restricted to citation forms; in practice, semantic factors tend to override syntactic factors in determining which words receive accent (Cutler and Isard, 1980; Ladd, 1980). Speakers adjust the relative prominence of the words they speak so as to communicate their message most efficiently. Moreover, this process is not necessarily a one-off assignment made at a relatively high-level utterance planning stage. The speaker can be shown to be monitoring prosody and adjusting it with the listener's comprehension in mind. This conclusion arises from some recent work on the way slips of the tongue, once made and detected by a speaker, are corrected.

Of course, not all slips of the tongue are detected by the speaker; and not all slips which are detected are corrected. When a correction is issued, however, the correction may have the same prosody as the original utterance, or it may be given a very different prosodic contour. This seems, on the face of it, to be a trivially true observation; but it is not trivial. The continuum of prosodic divergence between original utterance and correction is bimodal, not continuous. Goffman (1981) first noticed this phenomenon in radio announcers; some corrections hardly interrupt the flow of speech at all, and in particular the prosodic pattern is not altered at all, while others result in a radical change in the original prosody. Cutler (1983) took pitch and amplitude measurements of a large corpus of error corrections; each distribution of the difference between original and corrected utterance was clearly bimodal. For example, (1) is a typical 'unmarked' correction: the peak pitch reading for the error word *(Mike)* is 139 Hz, for the correction *(Martin)* also 139 Hz.

(1) and bowls the first ball to Mike — Martin Kent
(2) then he himself loses the chance, that is he risks the chance of dying

In (2), on the other hand, *loses* was again spoken with a peak pitch reading of 139 Hz, but *risks,* the correction word, reached a peak of 217 Hz, a 56% increase. (2) is a 'marked' correction.

Levelt and Cutler (1983) investigated the determinants of error correction marking, in a large corpus of corrections collected by Levelt (1983). Firstly, they found that marking was only applied to corrections of real errors, not to correction for appropriateness (such as replacing a correct but general word by a more specific alternative). They argued that marking a correction amounts to accenting it, in order to emphasize the contrast between the correction and the original, incorrect utterance. This claim was further strengthened by the finding that the likelihood of marking a correction was a function of the *degree* of contrast between error and repair. The corpus in question consisted of speakers' descriptions of routes through a pattern of coloured dots, and word substitution errors were chiefly of two kinds: errors of direction, in which polar opposites were confused (e.g. *left* for *right, up* for *down);* and errors of colour, in which one of the eleven colour names in the pattern set was substituted for another. Levelt and Cutler argued that the degree of contrast was higher in direction errors than in colour errors, so that there should be a significantly greater probability of correction being marked with direction errors than with colour errors. Indeed, 72% of direction errors in the corpus were followed by marked corrections, but less than 50% of colour errors, a statistically significant difference. Levelt and Cutler concluded that how an error is corrected is determined by how the speaker perceives the error to have affected successful communication of the intended message. The more the actual utterance is at variance with the intended, the more likely it is that the speaker will adjust the prosodic structure to highlight the correction, thus drawing the listener's attention to the desired message.

The acceptability of segmental distortion

Further evidence from studies of slips of the tongue and the way they are corrected supports the general claim that speakers' repair actions are determined by how much the slips are likely to have disrupted perception. For example, errors of lexical stress, such as (3)-(6), are corrected only rarely:

(3) you think its sarCASm, but it's not.
(4) . . . we're still enTHUSiastic.
(5) from my PROsodic — proSODic colleagues.
(6) everyone knows that ecoNOMists — that eCONomists . . .

In the corpus of errors and corrections analysed by Cutler (1983), it was found that correction correlated strongly with the effect of the stress shift on the vowel which would have been stressed had the word been uttered correctly. When that vowel was reduced in the error utterance, the error was corrected in 61% of the cases — as in (5) and (6), in which the target word's stressed syllable (the second syllable of *prosodic* and *economists* respectively) was spoken with a reduced vowel. When the target word's stressed vowel was given the same full vowel quality in the error as it would have received in correct production, however, a correction was issued only

21% of the time. (3) and (4) are examples — in neither case was the target word's stressed syllable (the first syllable of *sarcasm*, the fourth of *enthusiastic)* reduced, and in neither case was the error corrected. That is, speakers seem to be particularly concerned to correct lexical stress errors when they have resulted in gross distortion of phonetic segments, such as changing a full vowel to a schwa.

Distortion of phonetic segments can also result from the application of certain phonological rules of elision and assimilation in casual speech. Cooper and Paccia-Cooper (1980) have studied these effects intensively, particularly the extent to which speakers will apply assimilations and elisions across word boundaries. For example, palatalization is the rule which produces, from a [d] or a [t] followed by the glide [j], an affricate [dj] or [tf]; it can apply across word boundaries such as in 'did you' and similar phrases. Cooper and Paccia-Cooper examined the likelihood of palatalization applying across a word boundary as a function of the informativeness of the words preceding and following this boundary. For example, they varied the frequency of these words, comparing '. . . rode your horse . . .' with '. . . goad your horse . . . ', '. . . had utensils . . .' with '. . . had euglena . . .'. They found that varying the frequency of the [d] word (i.e. the word before the boundary) had absolutely no effect on the likelihood of speakers applying palatalization across the boundary; but varying the frequency of the [j] word had a dramatic effect: whereas with relatively high frequency phrases such as 'had utensils' over one-third of the productions were palatalized, with low frequency phrases such as 'had euglena' the frequency of palatalization dropped to 10%. The effect of contrastive stress was similar. Stressing the [d] word did not significantly inhibit palatalization; stressing the [j] word, however, almost completely suppressed it.

Cooper and Paccia-Cooper concluded that distorting the end of words does not concern the speaker greatly; speakers take pains, however, to avoid distorting the *onset* of words if the words are particularly informative (e.g. of low frequency; or contrastively stressed). It is difficult to conceive of an explanation for this effect in terms of demands of the production process alone. But again, the value to the perception process is quite obvious: the onsets of words are, for perception which takes place in time, the most crucial portions, and distortion of initial segments is likely to disrupt perception to a much greater extent than distortion of final segments.

Thus speakers' choices in on-line speech production — whether to correct a misplaced stress, whether to casually distort a word boundary — appear once again to be guided first and foremost by the requirements of their listeners' perceptual processes.

The curious case of word boundaries

The previous section discussed some circumstances under which casual speech processes may obscure word boundaries. In fact, word boundaries pose something of a problem for speech perception; in particular, they pose a problem for an account of speech production which invokes perceptual constraints.

Consider the current state of automatic speech recognition research. Isolated word recognizers are both feasible and available using current technology. Continuous speech recognizers, however, are simply still beyond current knowledge. The problem which so far has not been solved is that of segmentation. If speech recognizers could be supplied with reliable information about where each word in a continuous utterance began and ended, the successful construction of automatic continuous speech recognizers would be very close. But word boundary information is, at least, not reliably enough coded that speech scientists have as yet been able to make machines detect it.

The problem for the present argument is, therefore: despite all the examples cited above of speakers constraining their output in many and varied ways to make things easy for the listener, the one thing which speakers could do which would be particularly useful for listeners, namely provide precise information as to where one word ends and the next begins, they do not.

Why might it be particularly useful to know the boundaries of words? Strictly speaking, what is required is not boundaries between words (in the orthographic sense) but between whatever constitutes the units of lexical representation. Meaning must be represented in discrete units; it is impossible for listeners to carry around complete semantic representations for any sentence they might conceivably ever hear. The task of speech understanding consists of translating sound into meaning, i.e. locating the (discrete) lexical representations which correspond to the continuous stream of spoken sounds. If the listener knew exactly where the speech sounds representing one discrete meaning unit ended and those representing the next began, the task of locating lexical matches would be considerably facilitated. Why, given that speakers appear to strive to do so much else for listeners, do they not provide word boundary cues? Four possible answers, each logically distinct, suggest themselves:

(a) Speakers do in fact produce usable cues to word boundaries, although speech scientists and engineers have as yet failed to identify them.

(b) In the production process, constraints deriving from perceptual systems may apply only up to a certain level; word boundaries are obscured by automatic processes operating beyond that level.

(c) There is a trade-off with the constraints imposed by characteristics of the production system, such that provision of word boundary cues could only be achieved at considerable cost in effort to the speaker.

(d) There is a trade-off with independent constraints imposed by characteristics of the perceptual system, such that marking of word boundaries would conflict with application of other perceptually determined effects.

The tentative answer which will be suggested in the following sections does not, however, correspond exactly to any of (a)-(d). It will be argued that both (a) and (d) are partly correct; but that, more importantly, it may be necessary to revise our conception of what the lexical unit is, or at least what the code for accessing it is. Word boundaries may not be what speech engineers think they are.

Stress patterns and segmentation

The processing of prosody is a somewhat neglected area of speech recognition research. It will be argued here that the problematic area of speech segmentation is one in which attention to the possible contributions of prosodic information could bring considerable advances, both in understanding human perception and in guiding automatic recognition.

Recent cross-linguistic work on segmentation in speech understanding has shown that the apparent units of segmentation may differ for speakers of different languages (Cutler, Mehler, Norris and Segui, 1983, 1986): the syllable appears to function as a segmentation unit for speakers of French but not for speakers of English. The search for units of perception has long exercised Psycholinguistics, and this new evidence is rather disturbing, in that it suggests that aspects of the segmentation process may be language-specific, which in turn implies that the proper model of speech recognition may differ for different languages or speakers. This conclusion is disturbing simply because the aim of Psycholinguistics is to model the general case of language perception and production, independently of language-specific variations. The perceptual unit model does not readily constitute a general model if the units in question may be vastly different.

Cutler and Norris (in press) have suggested a possible alternative model which is couched in more general terms and offers a potentially language independent framework for segmentation. This model draws a distinction between strong syllables (those with full vowels) and weak syllables (with reduced vowels, such as schwa). The basic claim is that the segmentation process treats the two types of syllable differently. Each full vowel, together with its syllabic onset, if any, is treated as a potential word onset. Reduced syllables are treated as unlikely word onset points.

In a stress language, like English, in which not all vowels are full, this means that strong syllables will be segmented in a way weak syllables are not. To demonstrate this, Cutler and Norris used a task requiring detection of a word embedded in a larger string. CVCC words such as *melt* were converted into non-words by having a VC string appended to them, so that the final consonant of the embedded word would then function as the onset of a second syllable. The vowel in this second syllable could be either full or reduced. Thus *melt* appeared in *meltive* (with strong second syllable: [meltajv] or *meltesh* (with reduced second syllable: [meltef]). Subjects were required simply to listen to a string of such two syllable nonsense words and to respond whenever they detected one beginning with a real word. It was predicted that detection time would be longer for the *meltive* examples than for the *meltesh* since in the former case the second syllable, -rive, would be segmented and treated as a potential new word, thus disrupting the extraction of information from both the first and second syllables necessary for the successful detection of the embedded word *melt*. In *meltesh* the reduced vowel in the second syllable would not trigger the segmentation process, so detection of the word spread over both syllables should not be impeded.

Note that alternative word recognition models do not predict this difference. On a standard syllabification analysis the syllable boundary of both *meltive* and *meltesh* falls between the two medial consonants, so a syllabic segmentation unit model (e.g. Mehler, 1981; Segui, 1984) should predict that, because both strings will be segmented at the same place into two syllables, each string should make detection of the embedded word equally difficult. Similarly, a strictly left-to-right auditory word recognition model (e.g. Marslen-Wilson, 1980) should predict that the embedded word would be recognized as soon as it ended, irrespective of what sounds followed it; thus, again, such a model should predict no difference between the two conditions.

In fact, as predicted by the full vowel model, embedded words were detected significantly more slowly when they were followed by a full vowel than when they were followed by a reduced vowel. When the VC endings were edited off and the experiment rerun as a word detection task, no difference was found between the words which had had full vowels edited off and those which had had reduced vowels edited off. Thus the original difference was surely due to the nature of the following vowel.

Cutler and Norris argued that directing attention to strong syllables makes good perceptual sense. In a stress language, strong syllables are acoustically clearer than weak syllables. Moreover, as Huttenlocher (1984) has shown, strong syllables contain more phonetically useful information than weak syllables. Huttenlocher calculated the number of words potentially satisfied by a broad phonetic transcription such as 'stop-vowel-liquid-stop', and found that discarding information in weak syllables did not significantly increase the size of the set of words satisfied by a particular transcription, whereas discarding strong syllable information increased set size several-fold. Thus the phonetic content of strong syllables is more informative than that of weak syllables, as well as being more perceptible due to simple acoustic advantages of greater duration and intensity.

Moreover, there is independent evidence that English listeners treat strong syllables as potential word onsets. Taft (1984) found that listeners preferred to segment ambiguous bisyllabic strings at strong syllable onsets. For instance, whereas a one-word form was chosen more often for [letas], which could equally well be *lettuce* or *let us*, a two-word solution was chosen more often for [invests], which could be either *invests* or *in vests*.

Thus the occurrence of strong syllables appears to be an important factor in segmentation. The rate of occurrence of strong syllables is the crucial ingredient in linguistic rhythm, be it stree-based in one language, syllable-based in another. This suggests that a key to understanding segmentation procedures in continuous speech recognition may lie in the processing of rhythm.

Rhythm in speech perception and production

There is a good deal of diverse evidence that rhythm provides useful information in speech perception. The disruption of rhythm certainly disrupts performance on many perceptual tasks. Martin (1979), for example, found that either lengthening or shortening a single vowel in a recorded utterance could cause a perceptible momentary alteration in tempo and increase listeners' reaction time to detect phoneme targets. Meltzer, Martin, Mills, Imhoff and Zohar (1976) similarly found that phoneme targets which were slightly displaced from their position in normal speech were detected more slowly. Buxton (1984) found that adding or removing a syllable on a word preceding a phoneme target also increased detection time.

These results suggest that listeners process a regular rhythm in a rather active way, using it to make predictions about temporal patterns; when manipulations of the speech signal cause these predictions to be proven wrong, perception is temporarily disrupted. There is yet further evidence which shows listeners to be actively following prosodic continuity. Wingfield and Klein (1971) demonstrated that prosodic breaks over-ride syntactic breaks in click location tasks — that is, more clicks are falsely reported to have been heard at the prosodic boundary, indicating that prosodic boundary marking is the most salient. Darwin (1975) similarly showed that prosodic continuity over-rides semantic continuity in shadowing.

The predictive use of prosody in speech understanding was particularly obvious in experiments in which phoneme targets on acoustically identical words were responded to faster when the target-bearing word was preceded by a prosodic contour indicative of sentence accent occurring at the target word's position (Cutler, 1976). Thus the target *d* was detected more rapidly in (7), in which the target-bearing word *dirt* is accented, than in (8), in which the accent falls on *rug* — even when tape-splicing had ensured that the word *dirt* was acoustically identical in both (7) and (8).

(7) She managed to remove the dirt from the rug, but not the grass stains.

(8) She managed to remove the dirt from the rug, but not from their clothes.

Since the only difference in the part of the sentence preceding the target was the prosody with which it was uttered, prosody must have been the source of the response time difference. It was argued that listeners were extracting from the prosodic pattern predictive cues as to where accent would fall, with a view to directing particular attention to the location of accent. Follow-up studies further investigated the components of the prosodic pattern contributing to this effect. When pitch variation was removed, i.e. the sentences were monotonized, acoustically identical targets were still responded to faster in sentences like (7) than in sentences like (8) (Cutler and Darwin, 1981). Thus intonational variation was not a necessary component of the predictive accent effect. In later experiments, however, sentence rhythm was manipulated, such that by the use of digital techniques the waveform was stretched or compressed and the temporal pattern of (7) imposed on (8) and vice versa, with all other components of the original prosody being left intact. In this case, the response time difference disappeared, which suggests that rhythmic factors are at least making a very strong contribution to the predictive value of prosodic contours.

Thus there is considerable converging evidence that listeners make active use of rhythmic structure in speech perception. Moreover, there is evidence that speakers are concerned to impose a regular rhythmic structure on their utterances where possible. Again, this evidence comes from research on slips of the tongue. Some slips of the tongue result in an alteration of the rhythm of the intended utterance — for example, slips in which a syllable is added or deleted, or in which stress placement is shifted. Analysis of such slips shows that the erroneous utterances are significantly more often more regular in comparison to the intended utterances than less regular (Cutler, 1980b). For example in (9) a syllable has been omitted from the intended word *interlocutor* to give the non-word *interlocker* (stressed on the first syllable):

(9) what the speaker thinks his interlocker knows

The resulting utterance clearly has a more regular rhythmic beat than the intended utterance, in that there is a constant number of weak syllables between any two strong syllables, whereas the intended utterance would have displayed a more varied pattern. Such regularization appears to show an underlying pressure towards rhythmicity in speech production, which occasionally expresses itself in the production of an error. A pressure towards rhythmicity may well admit of an explanation purely in terms of the demands of speech production itself; but on the other hand, given the evidence summarized in this section, it also accords very well with the notion of a speech production device closely attuned to the demands of the perceptual process.

Speaking for segmentation

The evidence of the preceding section suggests that rhythmic continuity is of very great importance to speech perception, since listeners use it so actively. Rhythmic continuity may be the main reason why speakers do not provide simple word boundary cues such as perceptible pauses between words. Words — or rather, units of lexical representation — can be of very differing lengths, so that marking the boundaries between them in some such prosodically sensitive way as pausing would of necessity result in a rather irregular and hence unpredictable rhythm. In this sense overt marking of word boundaries, on the face of it a great service to the perception process, could conflict with other perceptual demands — in this case, the demand for rhythmic continuity and regularity.

On the other hand, the very high degree to which perception is sensitive to rhythmic factors suggests that rhythm may perhaps offer an answer to the problem of segmentation. The essence of rhythm is the rate of occurrence of strong syllables. Listeners are very adept at computing speech rhythm, and using it predictively. It is not unreasonable to suggest that they may also be able to exploit it to generate word boundary information.

However, the word boundary information which they could extract would not be directly isomorphous with orthographic word boundary marking. Rhythm leads the listener to strong syllables. The results of the experiments described above suggest that strong syllables are indeed segmented in a way that weak syllables are not. Thus it may be that strong syllables are effectively the boundaries for lexical units. In some languages, all words begin with strong syllables; but in free stress languages, some words begin with weak syllables. The present proposal would imply that words beginning with weak syllables may not be accessed from the mental lexicon in strictly left-to-right order, but rather via their strong syllables. This is a radical proposal in terms of current models of word perception and speech recognition, since it violates the widely held assumption of 'sequential isomorphism', i.e. that the order of processing directly reflects the order of input.

However, it should be noted that independent arguments against the sequential isomorphism assumption have been offered by MacKay (Chap 18); and Huttenlocher and Goodman (Chap 19) have argued that a strictly left-to-right model of word recognition such as that of Marslen-Wilson (1980) cannot account for all word recognition performance. Therefore this proposal is in fact in line with other recent work. Moreover, the present proposal offers a solution to the word boundary problem which is directly in line with the other evidence on the relation between perception and production summarized above. Speakers accommodate their output to listeners' needs at all levels of the production process, including formulation of the details of the articulatory programme. Just as at other levels, speakers give listeners what they need at the word boundary level — and what they need at that level is prosody.

Conclusion

The evidence summarized in the preceding sections therefore presents a satisfyingly coherent picture. Throughout the speech production process, the demands of the perception process are operative, constraining word formation choices, blocking elisions and assimilations which might interfere with word recognition, prompting corrections of slips of the tongue only when comprehension is likely to be impaired. Even an apparent glaring exception to the pattern of perceptual sensitivity in production appears not to be an exception after all: although boundaries between orthographic words are not reliably marked in the speech signal, studies of the processing of rhythm suggest that listeners use rhythmic information to segment the speech signal into lexical units. There appears to be a strong pressure towards regularity of rhythm in the production of English, and regularity of rhythm is apparently just what listeners rely on to segment English. Thus the perceptual process is well served by the production process in all respects.

This is of course not to deny at all that production-internal factors constrain production. It would be extraordinary were perceptual exigencies to influence the production process in ways that were directly inimical to the needs of production. Regularity of rhythm, for instance, has an obvious role in facilitating speech production; indeed, Shaffer (1982) has argued that rhythm has a general beneficial organising function in all skilled motor performance. The background picture against which the present arguments should be considered is rather one in which production and perception processes co-operate at all levels. In fact, with respect to rhythmic processing, it has been argued that production and perception share an underlying timing mechanism (Keele, Pokorny, Corcos and Ivry, 1985). Speech production and perception play so important a role in our life that it should be no surprise to find that the two processes co-exist in cooperation rather than in competition. The present evidence of how speaking accommodates itself to listening is just further confirmation of this happy reciprocity.

Acknowledgment

The unpublished experiments described in the final two sections were supported by a grant from British Telecom. Jeffrey Bloom wrote the programs for temporal manipulation of speech waveforms, and Steve Bartram ran the experiments. Special thanks to Dennis Norris for many useful discussions.

Note

[1] Some researchers have claimed that normal identification of morphologically complex words also involves separation of affix and base (e.g. Taft and Forster, 1975; MacKay, 1976), although others have claimed that complex words have unanalysed lexical entries (e.g. Butterworth, 1982). Even if bases and affixes are always processed separately, the understanding of a neologism cannot be exactly the same process as the recognition of a known word, for the simple reason that we do notice when something we have heard is a made-up word. Therefore a model of the normal process based on recognition of the separate parts must allow for an additional process of recognition of the combination.

References

Butterworth, B. (1982). Lexical representation. In B. Butterworth (Ed.) *Language Production, Vol. 2: Development, Writing and Other Language Processes.* London: Academic Press.

Buxton, H. (1984). *Rhythm and Stress in Speech.* Unpublished PhD Thesis, University of Cambridge.

Cooper, W. E. and Paccia-Cooper, J. (1980). *Syntax and Speech.* Cambridge, MA: Harvard University Press.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics,* 20, 55-60.

Cutler, A. (1980a). Productivity in word formation. *Papers from the Sixteenth Regional Meeting, Chicago Linguistic Society,* 45-51.

Cutler, A. (1980b). Syllable omission errors and isochrony. In H. W. Dechert and M. Raupach (Eds) *Temporal Variables in Speech.* The Hague: Mouton.

Cutler, A. (1981). Degrees of transparency in word formation. *Canadian Journal of Linguistics.,* 26, 73-7.

Cutler, A. (1983). Speakers' conceptions of the functions of prosody. In A. Cutler and D. R. Ladd (Eds) *Prosody: Models and Measurements.* Heidelberg: Springer.

Cutler, A. and Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, 29, 217-24.

Cutler, A. and Isard, S. D. (1980). The production of prosody. In B. Butterworth (Ed.) *Language Production.* London: Academic.

Cutler, A., Mehler, J., Norris, D. and Segui, J. (1983). A language-specific comprehension strategy. *Nature,* **304,** 159-60.

Cutler, A., Mehler, J., Norris, D. G. and Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language,* 25, 385-400.

Cutler, A. and Norris, D. G. (in press). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception & Performance.*

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. In A. Cohen and S. G. Nooteboom (Eds) *Structure and Process in Speech Perception.* Berlin: Springer.

Goffman, E. (1981). Radio talk. In E. Goffman (Ed.) *Forms of Talk.* Oxford: Blackwell.

Huttenlocher, D. P. (1984). *Acoustic-Phonetic and Lexical Constraints in Word Recognition: Lexical Access using Partial Information.* Unpublished M.Sc. thesis, MIT.

Keele, S. W., Pokorny, R. A., Corcos, D. M. and Ivry, R. (1985). Do perception and motor production share common timing mechanisms: a correlational analysis. *Acta Psychologica,* 60, 173-91.

Ladd, D. R. (1980). *The Structure of Intonational Meaning.* Bloomington: Indiana University Press.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition,* 14, 41-104.

Levelt, W. J. M. and Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics,* 2, 205-17.

MacKay, D. G. (1976). On the retrieval and lexical structure of verbs. *Journal of Verbal Learning and Verbal Behavior,* 15, 169-82.

Marslen-Wilson, W. D. (1980). Speech understanding as a psychological process. In J. C. Simon (Ed.) *Spoken Language Generation and Understanding.* Dordrecht: Reidel.

Martin, J. G. (1979). Rhythmic and segmental perception are not independent. *Journal of the Acoustical Society of America,* 65, 1286-97.

Mehler, J. (1981). The role of syllables in speech processing: infant and adult data. *Philosophical Transactions of the Royal Society,* B **295,** 333-52.

Meltzer, R. H., Martin, J. G., Mills, C. B., Imhoff, D. L. and Zohar, D. (1976). Reaction time to temporally displaced phoneme targets in continuous speech. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 277-90.

Segui, J. (1984). The syllable: a basic perceptual unit in speech processing? In H. Bouma and D. G. Bouwhuis (Eds) *Attention and Performance X: Control of Language Processes.* Hillsdale, NJ: Erlbaum.

Shaffer, L. H. (1982). Rhythm and timing in skill. *Psychological Review,* 89, 109-22.

Taft, L. (1984). *Prosodic Constraints and Lexical Parsing Strategies.* Ph.D. Thesis, University of Massachusetts.

Taft, M. and Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior,* 14, 638-47.

Valian, V. V. and Wales, R. J. (1976). What's what: talkers help listeners hear and understand by clarifying sentential relations. *Cognition, 4,* 115-76.

Wingfieid, A. and Klein, J. F. (1971). Syntactic structure and acoustic pattern in speech perception. *Perception and Psychophysics,* 9, 23-5.